Scan-Centric, Frequency-Based Method for Characterizing Peaks from Direct Injection FT-MS Experiments

Robert M Flight1,2,3, Joshua M Mitchell1,2,3,4, and Hunter NB Moseley1,2,3,4,5,✉

2022-03-29 17:13:33

1 Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA  
2 Department of Molecular & Cellular Biochemistry, University of Kentucky, Lexington, KY 40536, USA  
3 Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky, Lexington, KY 40536, USA  
4 Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA  
5 Department of Toxicology and Cancer Biology, University of Kentucky, Lexington, KY 40536, USA

✉ Correspondence: [Hunter NB Moseley <[hunter.moseley@uky.edu](mailto:hunter.moseley@uky.edu)>](mailto:hunter.moseley@uky.edu)

**Abstract:** We introduce a novel method for characterizing peaks from direct injection FT-MS experiments developed on metabolomics datasets that utilizes frequency values derived directly from the spacing of raw m/z points in spectral scans. Our method incorporates several aspects designed to make the derived intensities trustworthy, and suitable for use in our previously published peak assignment algorithm SMIRFE. Notably, the inclusion of an intensity independent noise removal, and normalization of scan level data results in a reduction of median relative standard deviation from **X** to **Y**, and much better fit of peak intensities to relative natural abundances of assigned isotopologues.

**Keywords**: Fourier-transform mass-spectrometry

## Introduction

What can we say about FT-MS data? There is more and more of it, and more of it applied to metabolomics. Profile mode, multi-scan direct injection FT-MS is an interesting approach because it allows us to handle cases where something might go wrong in an individual scan, while still extracting useful data.

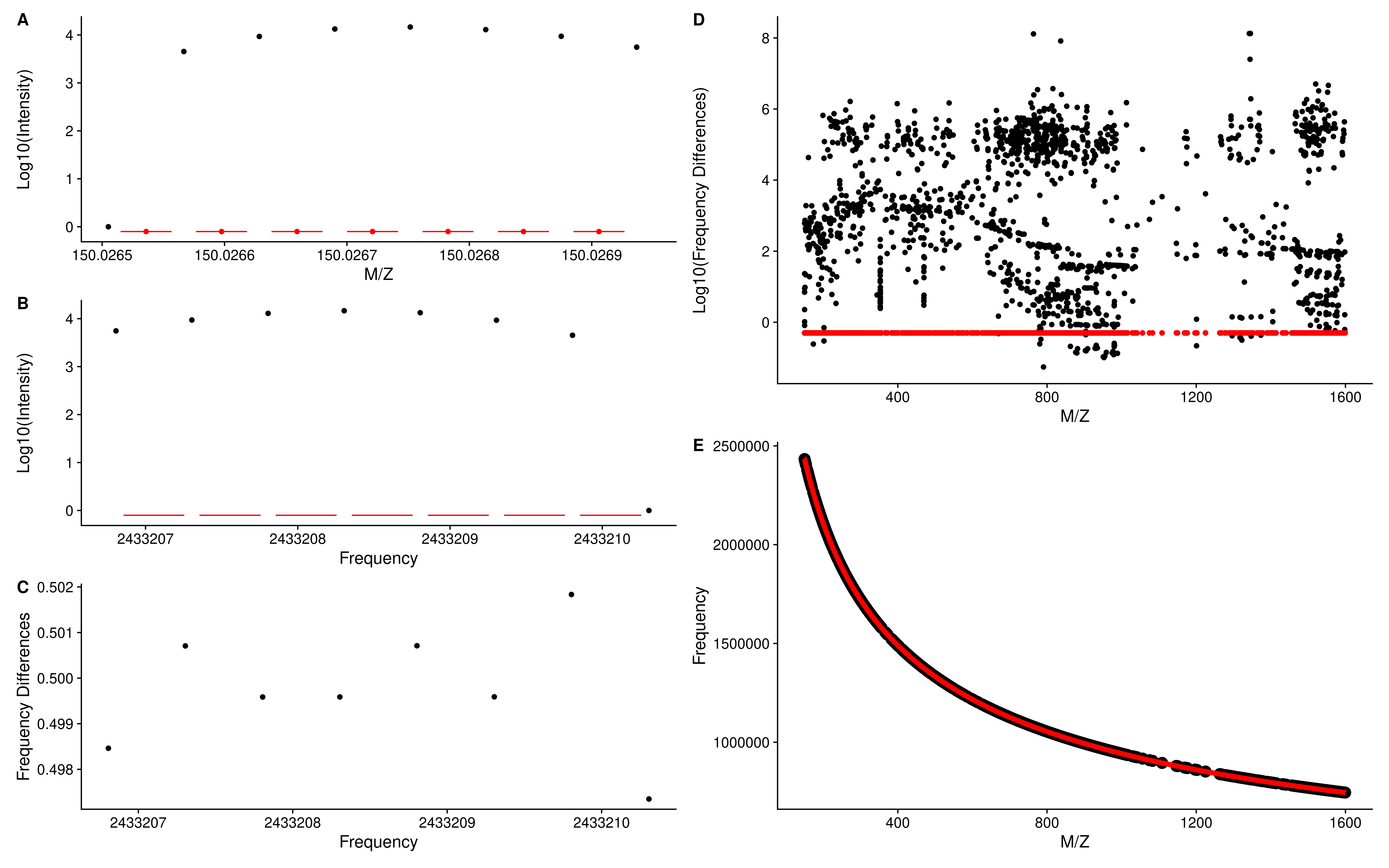
## Results

### Naively Averaged Data Have Bad Relative Intensities

To motivate our solution, we generated peak lists using the peak exporting functionality in Xcalibur (a process we have observed being used to generate datasets), as well as merging scans (see Methods) in the MSnbase R package. In all cases, for a known assigned set of peaks, the intensity ratios between the assigned peaks deviate from the intensity ratio predicted by their natural abundance probabilities (NAPs), as shown in Figure X, with some solutions giving different amounts of error. We do note that when all of the peaks are present in a single scan, their relative intensity ratios are much closer to the theoretical ratios based on NAP.

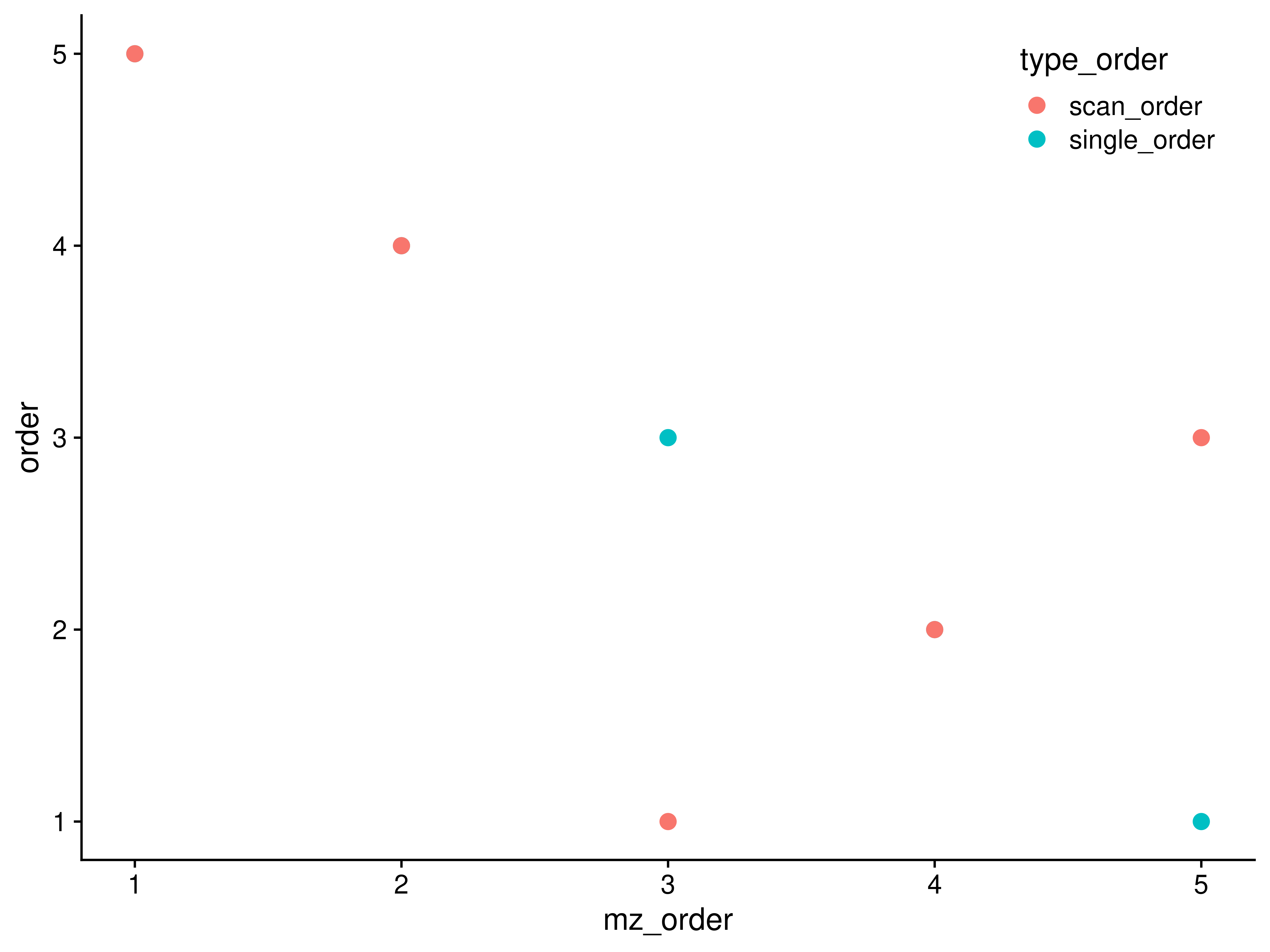
### m/z to Frequency

FT-MS data from the Thermo-Fisher Orbitrap instruments used to acquire the data does not provide any information about the raw frequency data. Outside of the meta-data, it merely contains the m/z and intensity values for profile spectra acquired across multiple scans. However, the frequency can be calculated by dividing the average m/z of two adjacent points by their difference (Figure 1A, red points representing the average m/z of two adjacent points, red lines representing the difference between the two adjacent points). The subsequent differences in frequency should be relatively constant with respect to m/z, in contrast to the differences of adjacent m/z points, as shown in Figure 1B and C. The Thermo Fusion instrument from which most of our collaborators data has been acquired, at a resolution of 400K, has a mode of 0.5, as shown in Figure 1C and D. Restricting to those points that fall into a narrow range of frequency differences (0.49 - 0.51), a model of frequency to m/z can be generated (see Methods), with an example shown in Figure 1E. This model seems to fit the known relationship between frequency and m/z, where the frequency is related to . This is useful, because some of the subsequent steps in our workflow use sliding and tiled windows where it is assumed that the sliding windows contain the same number of data points. The m/z point-to-point differences are not constant, but can be approximated by a lasso linear model; however, it is very difficult to create a lasso model with an intercept of 0. In addition, we would also need to vary the width of sliding windows according to the m/z difference at a particular m/z based on the lasso model. Frequency based points suffer none of these drawbacks, and the conversion from m/z can be derived from the raw profile level data itself, which is incredibly useful.



**Figure 1.** **A**: Intensity vs m/z for a single peak from a single scan. Red lines denote the differences between each point, and red dots the average between the pair of points. The difference over the point is used to derive the frequency values in **B**. **B** plots the intensity vs the converted frequency points derived from **A**. The red lines denote pairwise differences, which are shown in **C** for this single peak. The differences for all peaks across all scans vs m/z are shown in **D**, with those differences that lie within 0.49 - 0.51 shown in red. **E** shows the plot of derived frequency *vs* m/z, with fitted values from the linear model in red.

The m/z to frequency models are calculated for each scan, and the square root term from all scan level models are checked for outliers based on the interquartile ranges across all scans in a sample. While scan specific models **could** be used in the conversion of m/z to frequency, doing so results in changes to the relative peak ordering compared to m/z space, as shown in Figure 2. Therefore, a single model for all scans based on the scan with the slope closest to the median of slopes across scans is used for converting *all* remaining scan level data.

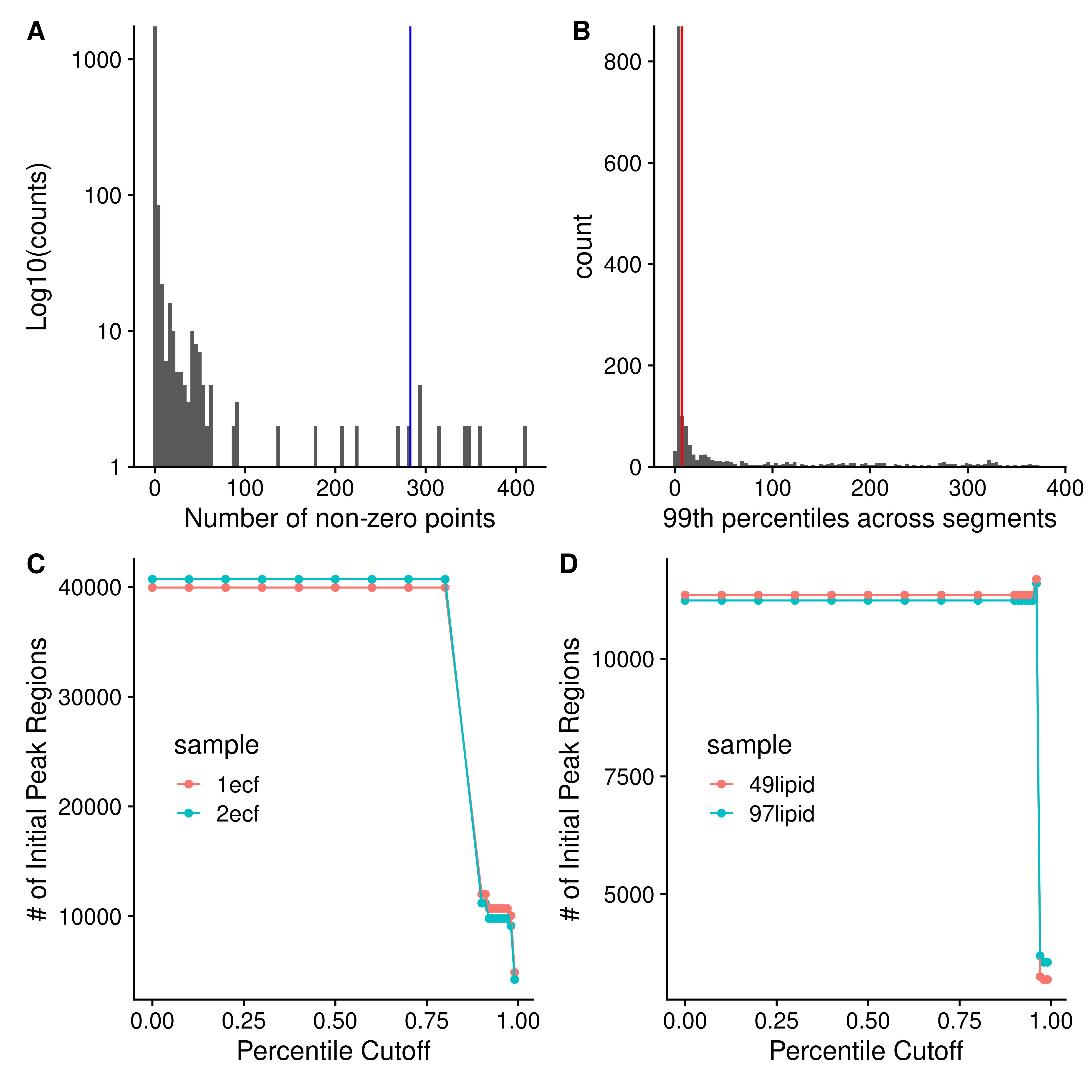


**Figure 2.** Peak ordering in m/z compared with ordering in frequency space when a single m/z to frequency model is used or scan specific m/z to frequency models are used. For a single peak, the scan level peak m/z’s were extracted, and then frequency values for those m/z generated using a single common model of m/z to frequency (*single\_order*), or models derived from each scan (*scan\_order*). A subset of the peaks end up out of order using scan specific models, implying that a single model should be used across all the scan level data.

Although the original model is created from only those points that had frequency point-to-point differences within a narrow range, **all** m/z points are converted to frequency for subsequent steps in the workflow.

### Sliding Window Density to Remove Noise

In a dataset of this nature, we expect that much of the data is really just noise, and doesn’t contribute that much to the analysis. Furthermore, it is expected that noise is randomly distributed across the scans. Therefore, if we slide a window across the data and sum the number of non-zero points in each window, we expect that most of the data we encounter is actually noise. Subsequently, we divide the counts into tiled regions of a set size (2000 in this work), and examine the 99th percentile of the sliding window counts within the tiled region. A histogram of sliding window counts in a single tiled region is shown in Figure 3A. The distribution of 99th percentiles across all of the tiled regions for a single sample is shown in Figure 3B. We can use the median (50th percentile) value of all the 99th percentiles from all of the tiled regions to determine a cutoff value above which any non-zero points are potentially **signal** and not just noise. In the example below, the median of the tiled windows is 5, resulting in a cutoff of 7.5. The number of peaks detected without applying a cutoff explodes anywhere from 3-fold (from 3,552 to 11,237 in the 97lipid sample), to 10-fold (from 4,227 to 40,701 in the 2ecf sample) with a vast majority of the peaks being present in a single scan. By discarding those windows with a non-zero point density below 1.5x the median of 99th percentiles, and merging the remaining adjacent windows, the regions most likely to contain actual, informative peaks are rapidly and easily determined.

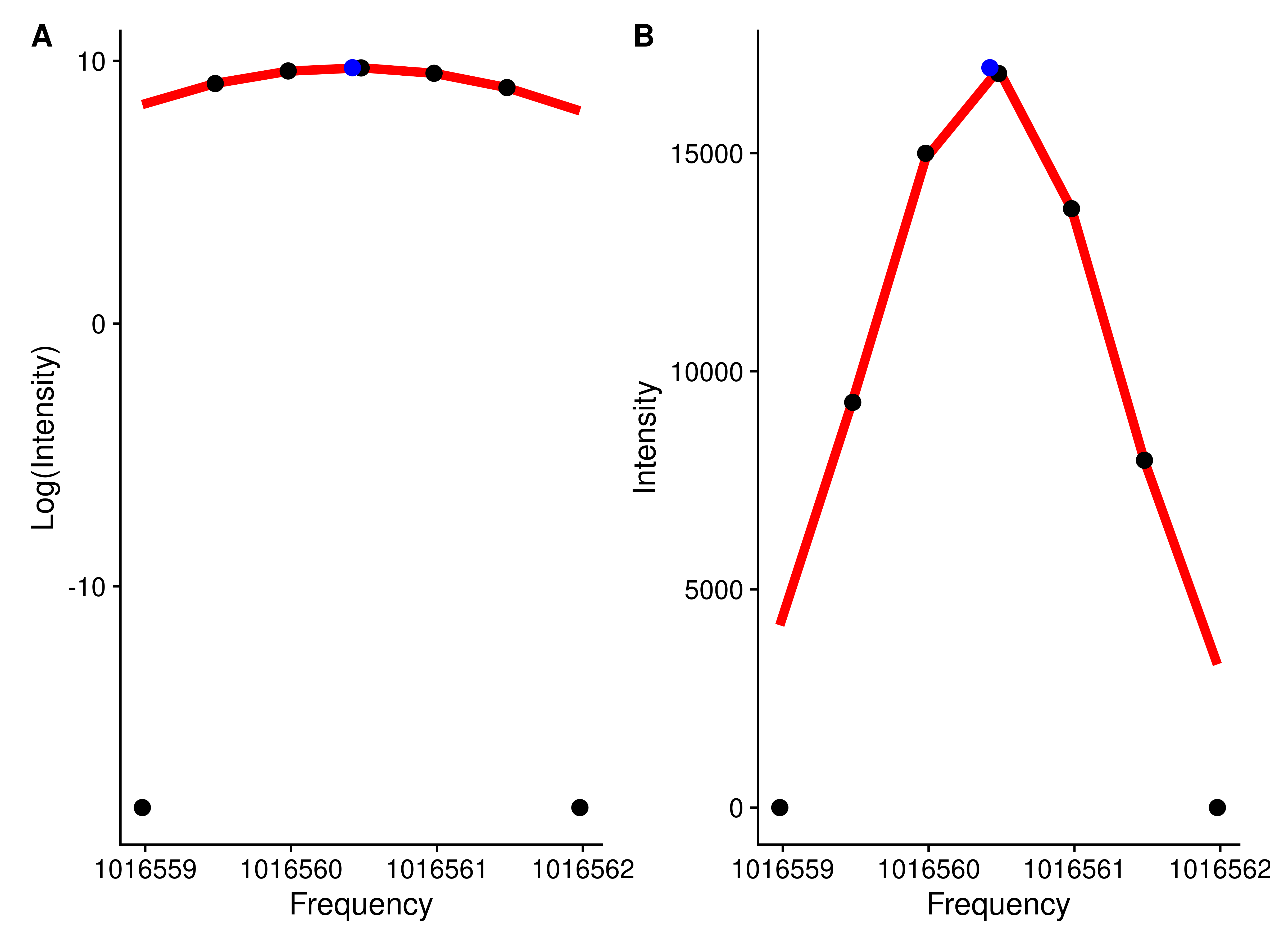


**Figure 3.** **A**: Histogram of the number of non-zero points from sliding windows ten frequency points wide and a stride of one point, from a single tiled window 2000 points wide. Blue line is the 99th percentile. **B**: Histogram of the 99th percentile cutoffs from all of the tiled windows. Red line is the median x 1.5. **C** and **D** show the number of initial regions as a function of the percentile cutoff used for either the AA ECF (C) or lipid (D) samples.

For each region initially created, the peaks in each scan within that region can be characterized using a quadratic fit of log(intensity) to m/z.

### Peak Characterization Using Quadratic Fit

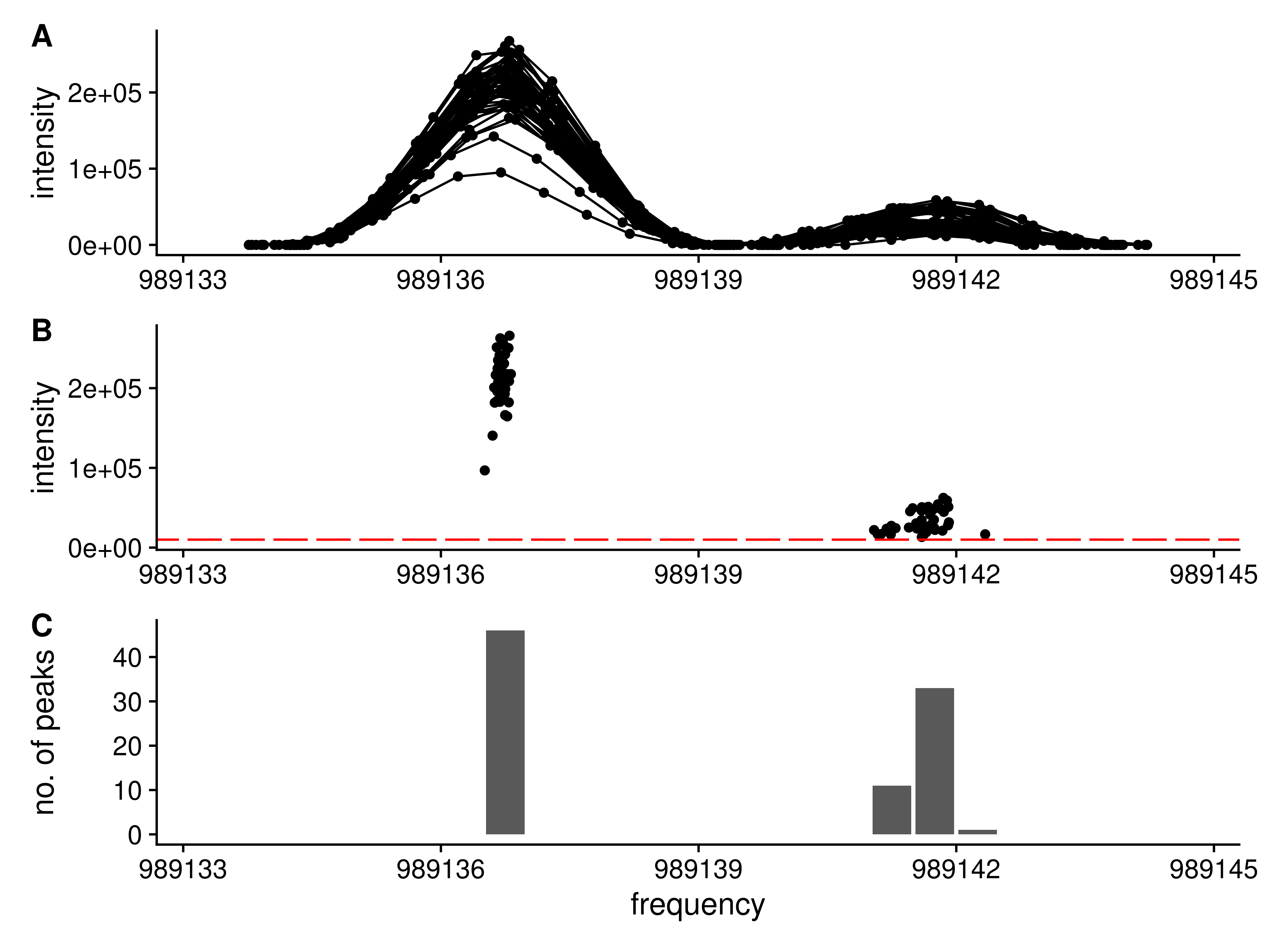
Although many other types of mass-spectrometry data suffer from a variable and noisy baseline, the scan-level profile data from the Thermo Fusion has a baseline of 0 due to manipulations in the Thermo firmware, making the determination of the centroided values considerably easier. For the peak characterization (centroiding), we use a simple quadratic model of log(intensity) to m/z. We add a small constant to enable using the zero values, and weight the values by their ratio to the most intense value, which is normally the value closest to the center of peak, helping to ensure that the **true** centroid is determined. From the fitted model, we can derive the centroided center and the intensity of the peak, as shown in Figure 4.



**Figure 4.** Log-intensity (A) and Intensity (B) of points for a single peak against frequency. Black points are the original data points, and the blue point represents the calculated centroid.

### Breaking Up Initial Regions

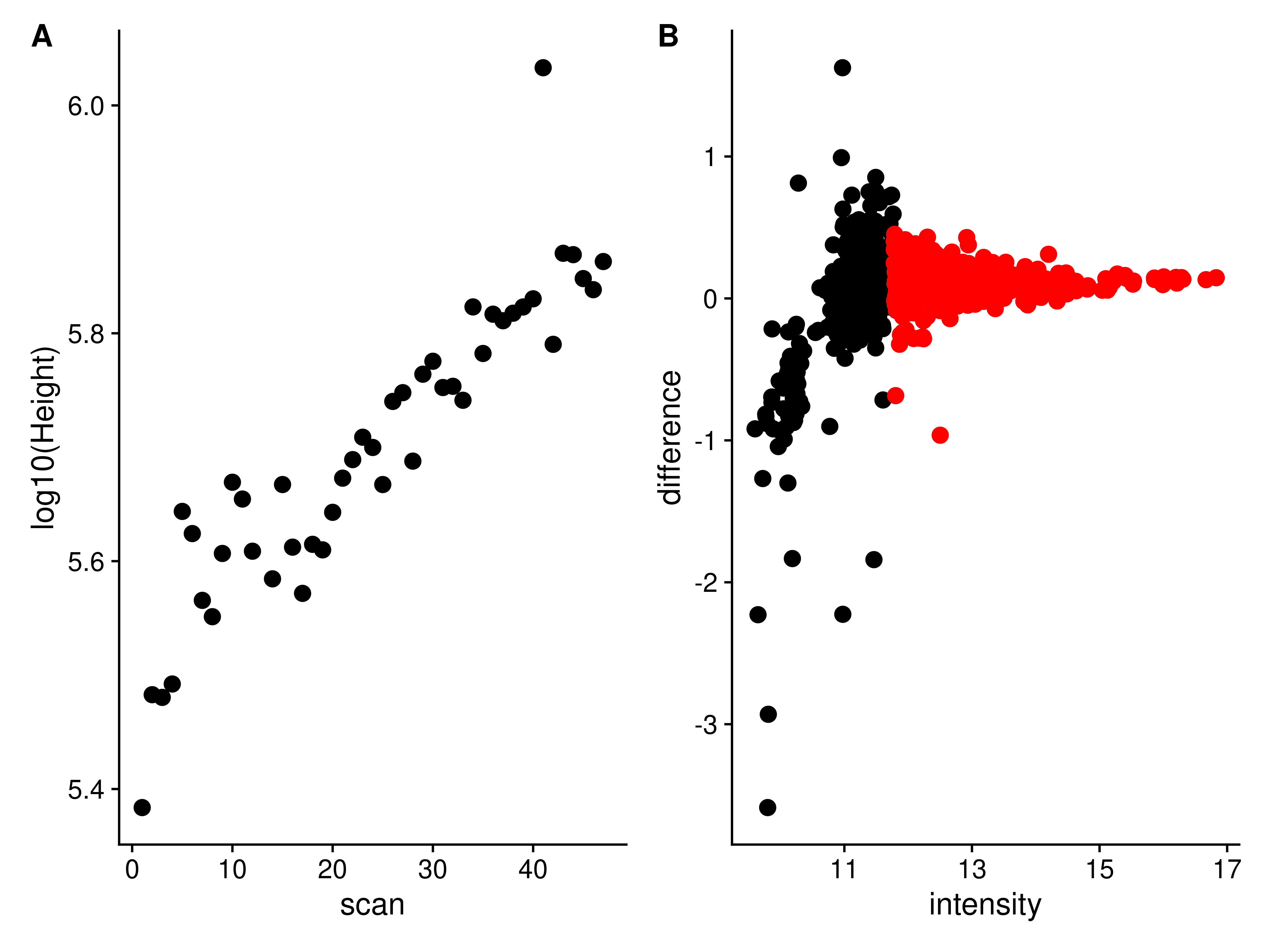
With the characterized (centroided) peak data from across scans within each region, it is then important to determine if only one or multiple “peaks” are actually present in the region. Our solution to this is to define breaks between **actual** peaks as a single frequency bin with zero characterized peaks within it. The frequency bins are created from tiled windows that are one frequency point difference wide. Adjacent non-zero frequency bins are merged to comprise a single peak region. Figure 5 shows an example where an initial region is broken up into two regions based on the characterized peak centers.



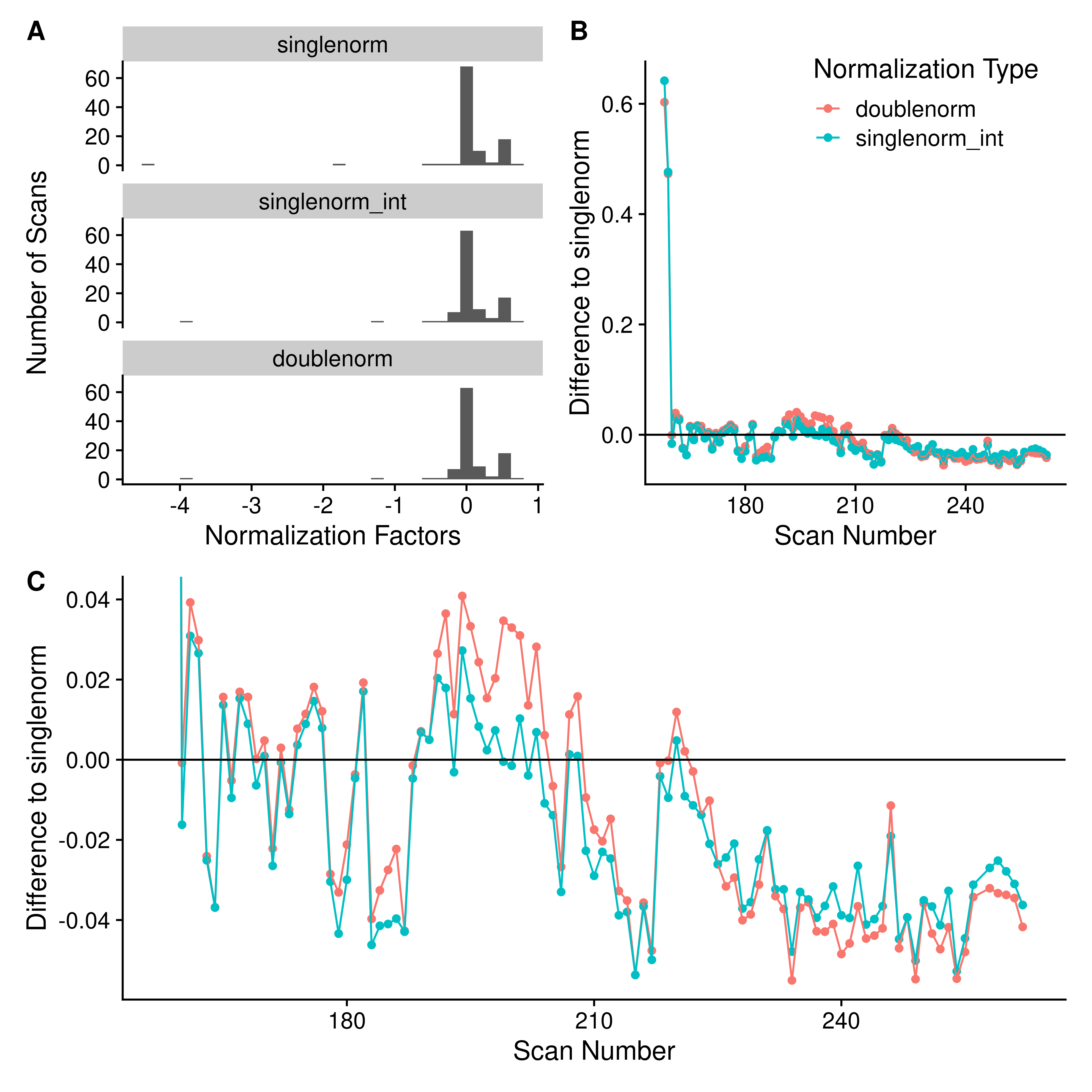
**Figure 5.** Splitting a single region into two regions based on the peaks that are present. **A**: The full set of raw frequency and intensity data across all scans for the region are shown. Clearly the region has two separate peaks within it. **B**: The peak centroids (frequency and intensity) for each peak in black. The tiled regions (red) are used to quantify the number of peaks. **C**: The number of peaks within each tiled region shown as a histogram. Each group of non-zero adjacent regions will be merged to form a new peak region.

### Normalization of Scans

Due to differences in how many ions are captured in the trap and the limited dynamic range of the detector, the observed peak intensities for the same analyte vary slightly between scans (see Figure 5A and B for example). Using the median peak differences between scans, it is possible to normalize the peak intensities across scans. However, there are two issues with these peak intensities across scans: 1) some peak intensities correlate with the scan number (i.e. scan acquisition order); and 2) some peak differences between scans are correlated with intensity. The solution to **1** is to do a two pass normalization. After the first pass, the peaks whose intensity is correlated with scan order are detected. In the second pass, the correlated peaks are removed, and normalization is carried out again. Figure 6A contains an example peak whose intensity across scans is correlated with scan number. The solution to **2** (peak differences correlated with intensity) is to only use the most intense peaks, as shown in Figure 6B. The highlighted peaks in Figure 6B are those with an intensity greater than 0.7 of the maximum intensity observed in that scan, and at least visually, their differences are **not** correlated with intensity. If *all* peaks are used for normalization, a very different set of normalization factors will be generated than by using only the *most intense* peaks, as shown in Figure 7A and B.



**Figure 6.** **A**: An example of a peak whose height across scans is correlated with scan number. **B**: The peak differences to the same peaks in a reference scan are plotted against peak height. Black: Peaks with a height < 0.7 of the maximum. Red: Peaks with a height >= 0.7 of the maximum.

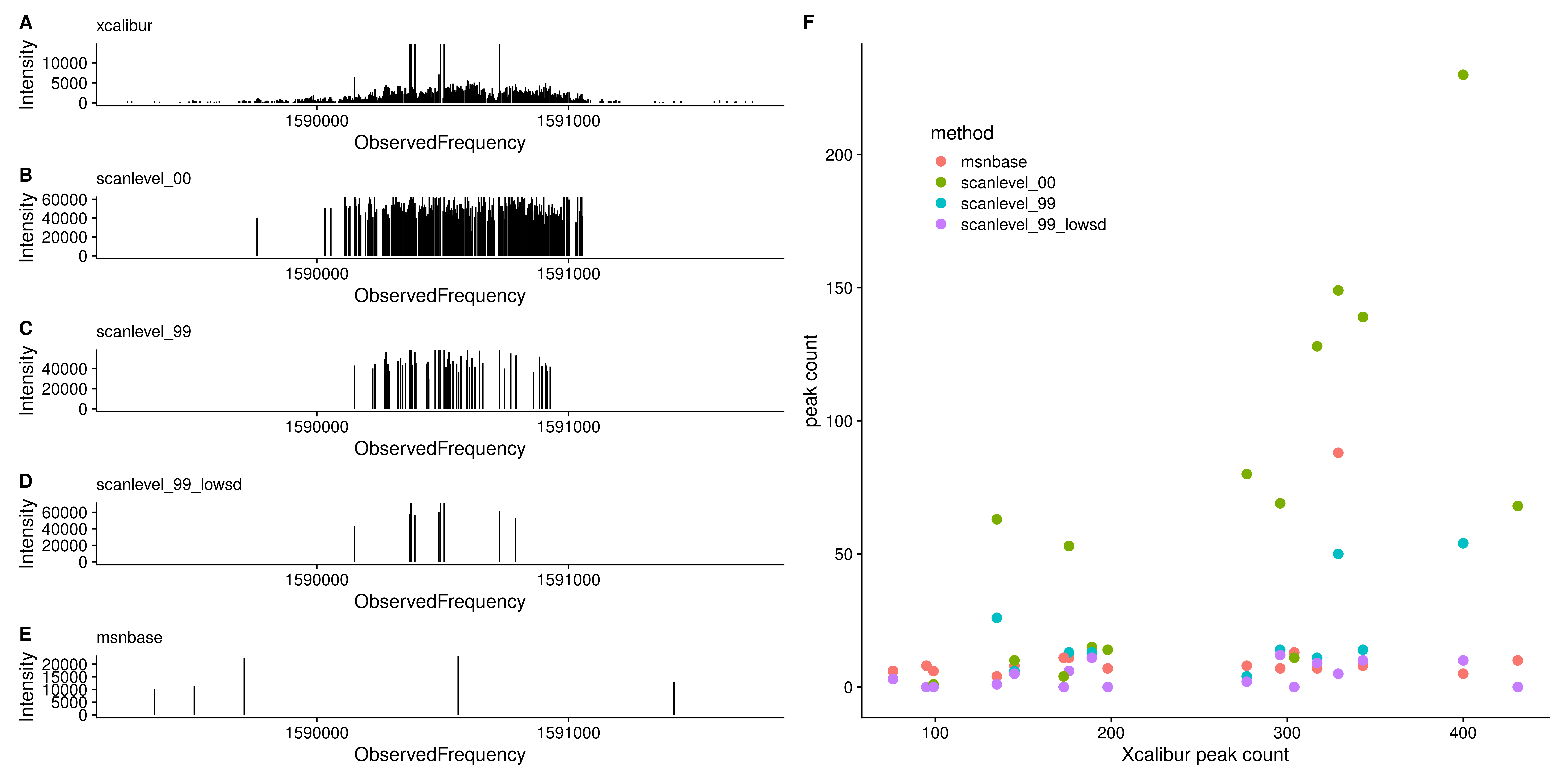


**Figure 7.** **A**: Histogram of scan normalization factors using either a single pass normalization using *all* peaks (singlenorm), single pass normalization using peaks with an intensity >= 0.7 of the maximum intensity (singlenorm\_int), or the double pass normalization removing peaks whose height is correlated with scan and using the most intense peaks (doublenorm). **B**: The difference of the normalization factors obtained from either doublenorm or singlenorm\_int compared to singlenorm.

### Removal of High Peak Density Artefacts

We have previously described the presence of high peak density (HPD) artefacts in FT-MS spectra [1]. Ideally, the peak characterization procedure should reduce their presence in the resultant reported peaks. Their presence should be minimized by removing noise peaks, and removing peak regions that have multiple reported peaks from the same scan. However, we expect they may also present as characterized peaks that have larger than expected frequency level standard deviations (FSD) when calculated across scans. These peaks can be detected by simply examining the distribution of FSDs and removing those that are outliers. To verify the removal of HPD peaks, we converted centroided m/z’s from XCalibur to frequency values using the previously calculated values for that sample. Peak density was measured using a sliding window ten points wide and a stride of one point.

Figure 8 shows a single HPD site detected in the ECF 2 sample, with the peaks from Xcalibur, as well as various scan-level processing and the peaks from centroiding using MSnbase. From this figure, we can see that the point density based noise filtering removes a large number of the peaks in the HPD site, while the frequency standard deviation removes further peaks that may be suspect. After removing the high FSD peaks, the number of peaks left in the HPD site are the same or less than those from MSnbase, and at least in this example, look more likely to be real peaks compared to those from MSnbase. Therefore, scan-level characterization allows us to keep what are likely **real** peaks in HPD sites, without consideration of peak intensity.



**Figure 8.** Comparison of HPD and high FSD sites in the ECF 2 sample. **A** - **E**:Peak plots for various peak processing modalities in a single HPD site. **A**:*xcalibur* peaks exported from Thermo Xcalibur after averaging scans. **B**:*scalevel\_00*: Scan-level peak characterization without any density based filtering (see noperc\_nonorm in Methods). **C**:*scanlevel\_99*: Scan-level peak characterization using the default point density filtering (see filtersd in Methods). **D**:*scanlevel\_99\_lowsd*: Same as *scanlevel\_99*, but removing any peaks marked as having a high frequency standard deviation. **E**:*msnbase*: peak centroids generated from *MSnbase*. **F**: Scatterplot of peak counts across all of the HPDs detected in the ECF 2 sample against the Xcalibur peak counts.

### Changes in Relative Standard Deviation (RSD)

Each step in the peak characterization either changes the overall number of peaks coming from each scan (sliding windows and breaking initial regions) or the overall intensity of the points within a scan. Therefore, one way to quantify any potential *improvements* in the characterized peaks is to look at the relative standard deviation (RSD) for the characterized scan level peak intensities (calculated as the standard deviation of peak heights across scans divided by the mean peak height), and compare them as each processing step is introduced.

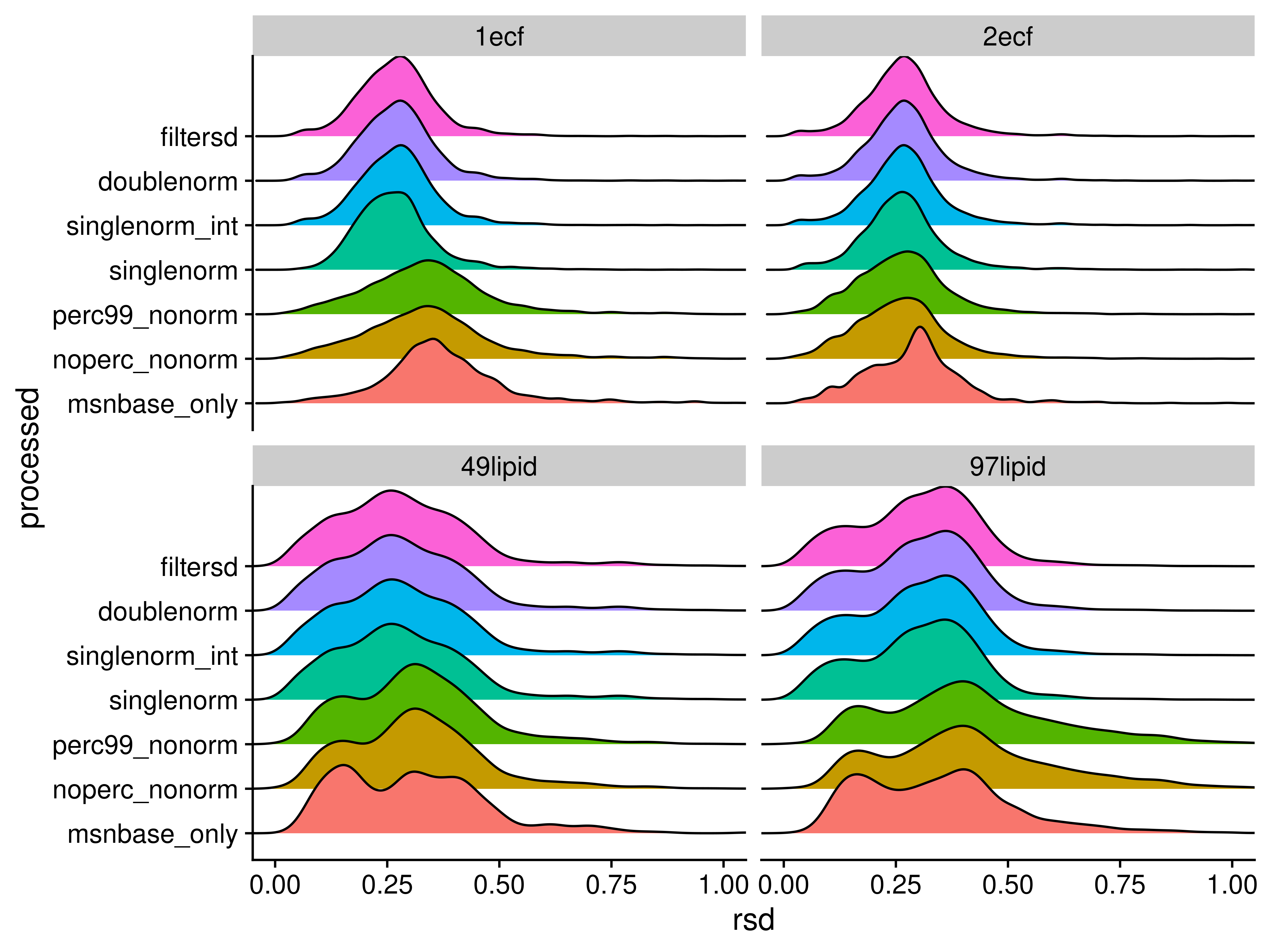


Figure 9. Density plots of relative standard deviations (RSD) of peak heights across scans for each of the processing methods. A peak had to be present in at least three scans for the RSD value to be reported.

Table 1. Relative standard deviation means, medians, modes and maximum observed values for each sample with different overall processing.

| processed | sample | mean | median | mode | max |
| --- | --- | --- | --- | --- | --- |
| doublenorm | 1ecf | 0.27 | 0.27 | 0.28 | 1.41 |
| filtersd | 1ecf | 0.27 | 0.27 | 0.28 | 1.41 |
| msnbase\_only | 1ecf | 0.37 | 0.36 | 0.35 | 1.19 |
| noperc\_nonorm | 1ecf | 0.34 | 0.33 | 0.34 | 1.54 |
| perc99\_nonorm | 1ecf | 0.34 | 0.33 | 0.34 | 1.54 |
| singlenorm | 1ecf | 0.27 | 0.26 | 0.28 | 2.19 |
| singlenorm\_int | 1ecf | 0.27 | 0.27 | 0.28 | 1.42 |
| doublenorm | 2ecf | 0.27 | 0.27 | 0.27 | 1.05 |
| filtersd | 2ecf | 0.27 | 0.27 | 0.27 | 1.05 |
| msnbase\_only | 2ecf | 0.29 | 0.29 | 0.30 | 0.99 |
| noperc\_nonorm | 2ecf | 0.26 | 0.26 | 0.28 | 1.06 |
| perc99\_nonorm | 2ecf | 0.27 | 0.26 | 0.28 | 1.06 |
| singlenorm | 2ecf | 0.27 | 0.26 | 0.26 | 1.03 |
| singlenorm\_int | 2ecf | 0.27 | 0.27 | 0.27 | 1.05 |
| doublenorm | 49lipid | 0.28 | 0.27 | 0.26 | 1.13 |
| filtersd | 49lipid | 0.28 | 0.27 | 0.26 | 1.13 |
| msnbase\_only | 49lipid | 0.31 | 0.30 | 0.15 | 1.14 |
| noperc\_nonorm | 49lipid | 0.31 | 0.31 | 0.31 | 1.13 |
| perc99\_nonorm | 49lipid | 0.31 | 0.31 | 0.31 | 1.13 |
| singlenorm | 49lipid | 0.29 | 0.27 | 0.26 | 1.13 |
| singlenorm\_int | 49lipid | 0.29 | 0.27 | 0.26 | 1.13 |
| doublenorm | 97lipid | 0.30 | 0.31 | 0.36 | 2.14 |
| filtersd | 97lipid | 0.30 | 0.31 | 0.36 | 2.14 |
| msnbase\_only | 97lipid | 0.35 | 0.34 | 0.40 | 2.14 |
| noperc\_nonorm | 97lipid | 0.42 | 0.40 | 0.40 | 2.11 |
| perc99\_nonorm | 97lipid | 0.42 | 0.40 | 0.40 | 2.11 |
| singlenorm | 97lipid | 0.30 | 0.31 | 0.36 | 2.14 |
| singlenorm\_int | 97lipid | 0.30 | 0.31 | 0.36 | 2.14 |

### Difference to Relative Natural Abundance

As an alternative to RSD, we can also compare the fit of relative intensities after assignment using SMIRFE [reference] to the theoretical relative natural abundances (relNAP) of the assigned isotopic molecular formula’s (IMFs) within the assigned elemental molecular formula’s (EMFs). Theoretically, we expect that lower quality data will have both lower numbers of assignments, and that for those things that are assigned, the fit between relative intensity and relNAP will be worse. To compare relative NAP to relative abundances, we only examined the assignments from the two samples containing ECF derivatized amino-acids, as we can limit the assignments to those that match expected derivatizations of the known amino-acids (see Supplemental for the expected EMFs, and expected relative NAPs for the individual IMFs).

Figure 10 compares the peak - peak natural abundance and height log-ratio differences generated using heights from Xcalibur and from our scan-level peak characterization.

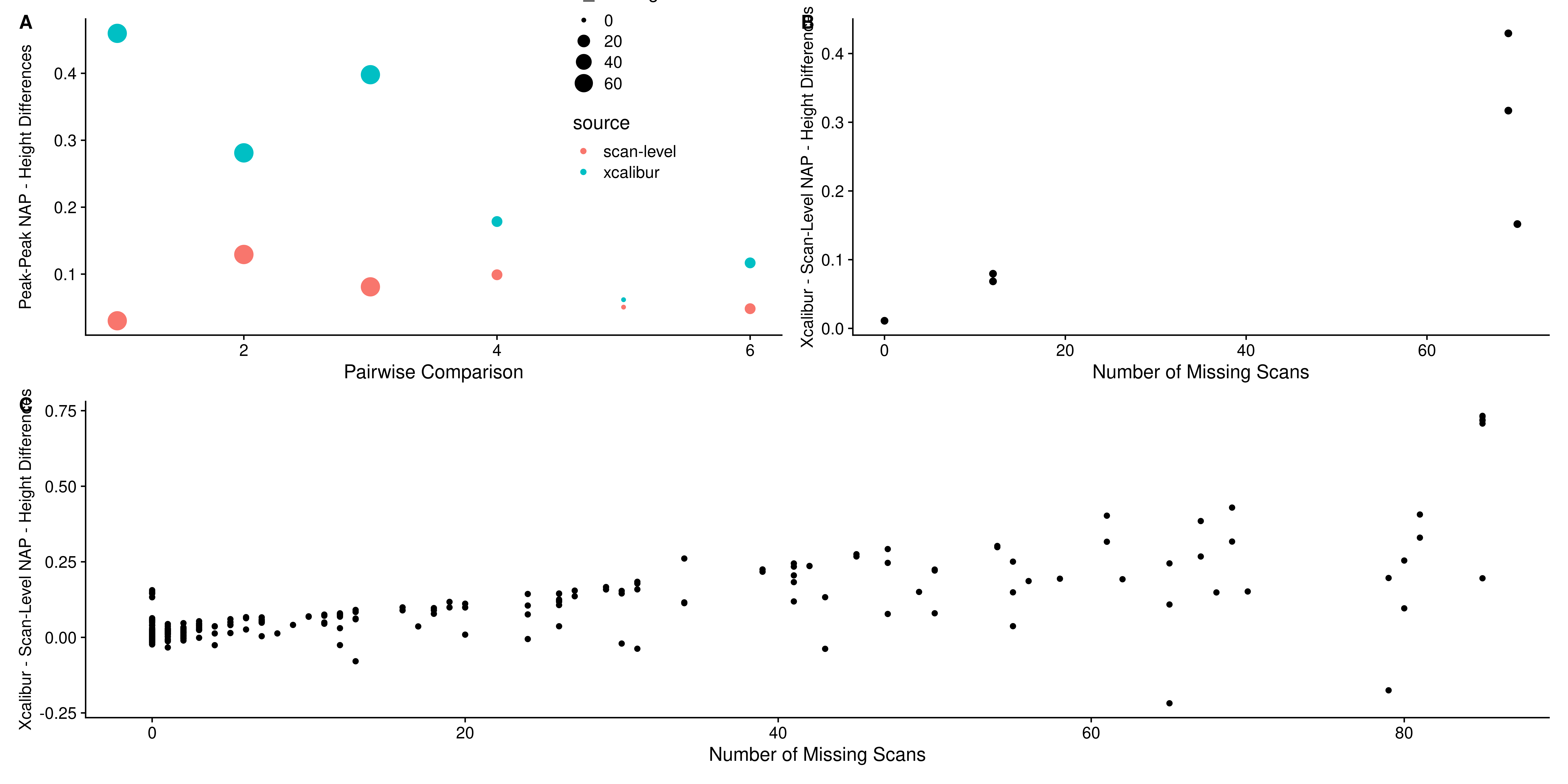


Figure 10. **A**. Peak - peak NAP - intensity differences from scan-level peak heights (red) and Xcalibur peak heights (blue) from the threonine amino acid assignments with Na adduct, with point size reflecting how many peaks were missing across scans. **B**. The difference of Xcalibur to scan-level ratios plotted directly as a function of the number of scans the peak was not found in. **C**. The differences of Xcalibur to scan-level ratios for all of the amino acid assignments in EMFs with more than a single peak in both ECF samples.

### Changes in Variance Across a Large Dataset

For large datasets, we expect that more correct peak intensities will result in better agreement between sample normalized peak intensities within a sample class. If incorrect peak intensities are being reported, then we would expect between class to within class variances to be very random, resulting in F-statistics close to 1.

## Discussion

## Materials and Methods

### Conversion of m/z to Frequency

The data consists of profile mode m/z spectra from multiple scans encoded as m/z and intensity values for each scan. No information about the original observed frequency values is available in either the raw files or the mzML files. However, proxy frequency values can be generated by averaging the m/z of adjacent points and dividing them by the difference. Ideally, the difference between subsequent points in this proxy frequency space is constant, in practice there is a range of differences in frequency space. The *actual, constant* difference can be obtained by examining the median of the frequency differences, and then constraining *useful* points (those that can be used for generating a model of frequency to m/z) to be within 2% of the mode value. These *useful* points can be used to construct a linear model relating m/z to frequency for individual scans based on the formula:

From the known physical properties of the Orbitrap, only the square root term should be necessary [2]. Practically, we have found the combination of square and cube-roots to provide a better fit, likely due to issues with slight imperfections in the orbitrap geometry, contributions from space charge effects and magnetronic motion, control of the magnetic fields, and the Fourier-like transform conversion used by the spectrometer. A frequency model is generated for each scan, and then a single model using the scan with the square-root term closest to the median of the square-root terms from all scans. We observed that this single model better preserved the relative ordering of the peaks in both m/z and frequency-space compared to the scan specific models (see Results). A single model rather than scan specific models is used for conversion because this was found to maintain the relative ordering of the peaks in m/z and frequency space (see Results).

To convert m/z back into frequency, we can use a similar model without the roots.

### Frequency Intervals

Two types of frequency intervals are used, sliding and tiled windows. In this work, the sliding windows are 10 points wide with a stride of one point. The tiled windows are one point wide with a stride of one point.

### Interval Range Based Data

To enable interval algebra, the frequency points are converted to single width intervals by multiplying (by a multiplier of 400 in this work), rounding to the nearest integer, and storing them as IRanges objects from the IRanges Bioconductor package [3]. The sliding and tiled windows are also converted to IRanges objects using this process.

### Peak Containing Intervals

To find intervals that contain points that represent actual signal and not just random noise, the number of non-zero intensity points in each sliding window are counted. Subsequently, we break these counts into fixed width tiles (default width of 2000) and calculate the 99th percentile of non-zero points for each tile. The median value x 1.5 of these 99th percentile values from the fixed width tiles is used as the cutoff value to determine which of the initial sliding regions should be kept as regions containing potential signal. Any sliding window with a non-zero count less than or equal to the cutoff value is removed, and the remaining sliding windows are reduced, where overlapping windows are merged to create the initial peak regions. The presence of zero intensity points are due to flooring implemented by the spectrometer when local spectral quality falls below a certain threshold. When this flooring is unstable and incomplete, a partial ringing phenomenom is observed [1].

Within each initial interval region, peaks in each scan are detected (see **Peak Detection**), and their centers are binned by the tiled windows. Adjacent tiled windows with non-zero peak counts are merged together, and any zero peak count tiled windows split the initial region into multiple peak interval regions. These interval regions should contain a single **real** peak that was detected in one or more scans.

### Peak Detection and Centroided Values

On a single scan level, possible peaks are detected by simple bump-hunting for two increasing points followed by two decreasing points using the find\_peaks function in the pracma package [4]. These possible peaks are then characterized using a weighted parabolic fit of log-intensity to position (where position is either m/z or frequency), and the weights for each point are the relative log-intensity compared to the maximum log-intensity for the peak.

From this weighted parabolic fit, the center, intensity, integrated area and sum-of-square residuals can be extracted for the peak. These characteristics are equivalent to the centroided peak center and intensity.

Before further processing, the regions are verified to have only one peak from each scan. If a scan has two or more peaks, then the scan level data in that region is discarded. Any regions that subsequently contain zero peaks are removed.

### Scan to Scan Normalization

Scans are normalized to a single *reference* scan based on the log-intensity differences of a subset of peaks present in at least the same number of scans as the 95th percentile of scan counts for the peaks. In addition, only those peaks with an intensity greater than 0.7 times the highest intensity peak in the scan are used. Pairwise scan-to-scan distances are calculated by taking the cartesian distances between log peak intensities present in both scans, and then the cartesian distance is summed across the scan-to-scan distance to provide an overall difference of each scan to all other scans. The scan with the lowest summed overall distance is chosen as the *reference* scan, and normalization factors for each scan are calculated as the median log peak intensity differences in compared to the reference scan. This normalization is done twice, once using all possible peaks, after which the correlation of peak intensity with scan order is checked, and those peaks with correlation of greater than 0.5 with scan order are removed, and the normalization factors are calculated again, and then applied to both the centroided peak height and the raw point intensities. Peaks correlated to scan order represent an artifact that we speculate results from a gradient in the sample well.

### Full Characterization

The full set of raw data points for each peak in each scan within a region is known based on the previously detected peaks. Therefore, the non-zero intensity, normalized raw data points across scans can be combined, and then characterized again using the weighted parabolic fit previously described. In addition to the data from the full set of raw points, the means and the standard deviations of the peak height and location can be derived from the scan-level peak characteristics previously calculated.

In addition to these values, the frequency point-to-point median difference is calculated across all of the raw data points for those points that could be used for modeling frequency to m/z, and this difference of a single point from the peak center is calculated in frequency space, and converted to m/z space to provide an “offset” value that is useful to define the search space around the peak for any assignment algorithm.

### Correction of Height and Standard Deviation

Ideally, each peak would be observed in every scan. However, some peaks are not observed in some scans due to the number of ions falling below the detection threshold. This results in a left-censored normal distribution of peak intensities across scans, which is expected for analytical measurements with detection limits [5]. To correct these, either a correction based on a model of the truncated normal distribution can be used, or the differences can be simulated by sampling from data that is present in most of the scans. To simulate the effect of peaks missing from some scans on the standard deviations, the peaks present in all scans are used. For each peak, a sample of the heights across scans are taken (ranging from 5% to 95% of scans), and a new standard deviation calculated for that fraction, and a ratio of the fractional standard deviation to the “true” standard deviation calculated. The ratio standard deviation across peaks can then be fitted to a cubic model of the fraction used, and a correction factor predicted for those peaks that are present in fewer scans. The corrected standard deviations can then be used to correct the mean height assuming that it is the result of a left-censored normal distribution [6,7].

### Frequency Standard Deviation Filtering

High-peak-density (HPD) artifacts [1] present as groups of singular peaks with higher than expected frequency standard deviations (FSDs) calculated from the scan-to-scan frequency peak locations. Outliers are detected by calculating the interquartile range (IQR) of the distribution of FSDs across the entire spectrum, and peaks with FSDs greater than the median plus 1.5 times the IQR (as implemented in boxplot.stats) are removed. The HPD detection algorithm from Mitchell et al. was re-implemented in R for this work to allow comparisons between it and the use of the FSD. For HPD detection, the peaks in excel output from Xcalibur were used after converting the m/z peak centers into frequency space. Sliding windows that are 1000 frequency points wide with a stride of 100 points were used for the density calculations.

### Multi-Class F-Statistics

To evaluate the F-statistic between multi-class samples, a much larger set of samples was used. In this case, 181 matched non-cancer and cancer samples previously used in the HPD detection manuscript [1]. All 181 samples were peak characterized using the most stringent method below as well as peaks generated from the MSnbase R package [ref], assigned using SMIRFE, and then peaks matched by shared EMFs across samples (see Supplemental materials for a description of peak matching and deciding the most likely EMF). For each matched peak across samples, we kept the peaks present in at least 25% of samples in both classes of sample (disease, non-disease), and calculated the F-statistic for peaks generated using scan-level peak-characterization and those from MSnbase.

### Samples and Overall Processing

Samples included two ethyl-chloroformate (ECF) derivatized amino-acid samples [reference] and two samples where lipid extracts were extracted from lung tumors [reference]. All four samples are positive mode, were converted from raw to mzML using msconvert from ProteoWizard (profile mode).

For each raw data file in profile mzML format, they were processed in these six ways:

1. No noise removal, no normalization, no frequency SD filtering (noperc\_nonorm)
2. Noise removal, no normalization, no frequency SD filtering (perc99\_nonorm)
3. Noise removal, single pass normalization with all peaks, no frequency SD filtering (singlenorm)
4. Noise removal, single pass normalization with high ratio peaks, no frequency SD filtering (singlenorm\_int)
5. Noise removal, two pass normalization, no frequency SD filtering (doublenorm)
6. Noise removal, two pass normalization, frequency SD filtering (filtersd)
7. Scan level peak centroids generated by MSnbase (using pickPeaks on the individual scans)
8. Scans merged and then centroids generated by MSnbase (using combineSpectra and pickPeaks)

For the two sets of MSnbase peaks, we also matched scan level peaks to merged peaks via two methods:

1. All scan level peaks within a 2ppm window are associated with the *merged* peak
2. Using the *filtersd* peaks, merged and scan level peaks are associated if they are within the start and stop m/z of the *filtersd* peaks

## Conclusions

## References

1. Mitchell JM, Flight RM, Wang QJ, Higashi RM, Fan TW-M, Lane AN, et al. New methods to identify high peak density artifacts in Fourier transform mass spectra and to mitigate their effects on high-throughput metabolomic data analysis. Metabolomics. 2018;14: 125. doi:[10.1007/s11306-018-1426-9](https://doi.org/10.1007/s11306-018-1426-9)

2. Ledford EB, Rempel DL, Gross ML. Space charge effects in Fourier transform mass spectrometry. II. Mass calibration. Anal Chem. 1984;56: 2744–2748. Available: <https://pubs.acs.org/doi/pdf/10.1021/ac00278a027>

3. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology. 2013;9: e1003118. doi:[10.1371/journal.pcbi.1003118](https://doi.org/10.1371/journal.pcbi.1003118)

4. Borchers HW. pracma: Practical numerical math functions. 2021. Available: <https://CRAN.R-project.org/package=pracma>

5. Flight RM, Bhatt PS, Moseley HN. Information-Content-Informed Kendall-tau Correlation: Utilizing Missing Values. 2022; 2022.02.24.481854. doi:[10.1101/2022.02.24.481854](https://doi.org/10.1101/2022.02.24.481854)

6. Truncated normal distribution. Wikipedia. 2022. Available: <https://en.wikipedia.org/w/index.php?title=Truncated_normal_distribution&oldid=1074943875>

7. Burkardt J. The Truncated Normal Distribution. : 35.